Comprehensive ID Mapping in Lustre 2.X

Stephen Simms (ssimms@iu.edu)
Joshua Walgenbach (jjw@iu.edu)

European Lustre Workshop
September 26-27, 2011

INDIANA UNIVERSITY

INDIANA UNIVERSITY

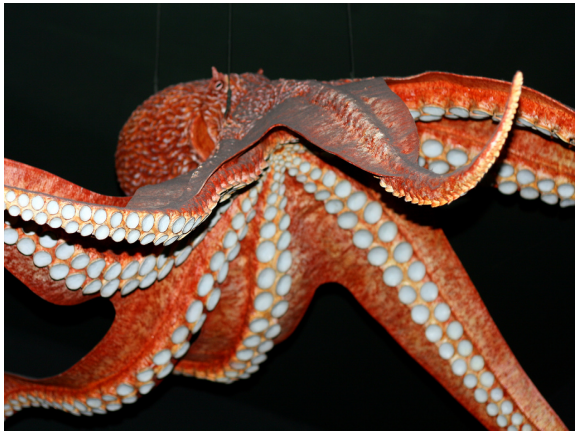
The Data Capacitor Project

NSF Funded in 2005

535 Terabytes Lustre storage

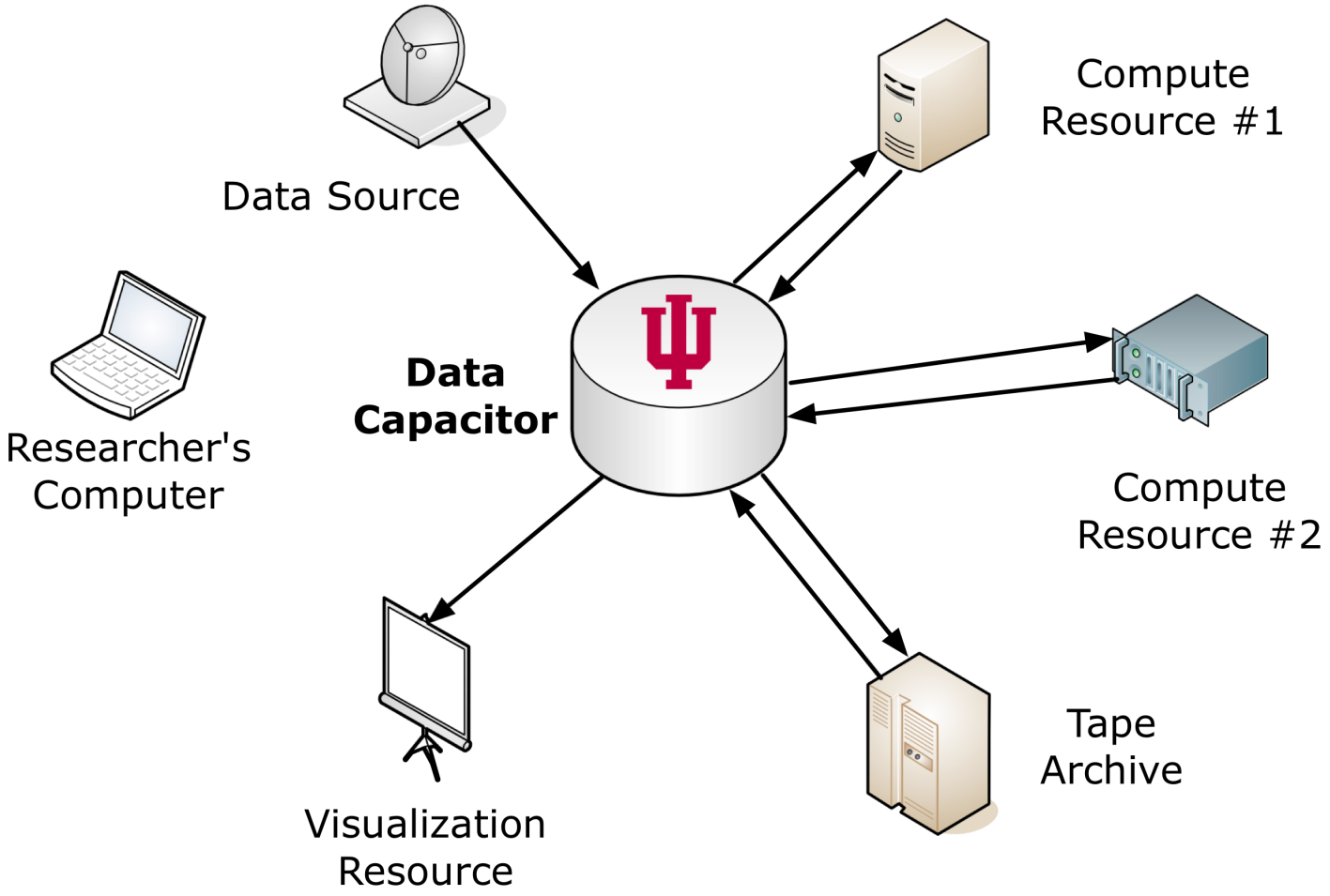
14.5 GB/s aggregate write

Short term storage



http://www.flickr.com/photos/shadowstorm/404158384/
http://www.flickr.com/photos/dvd5/163647219/
http://www.flickr.com/photos/vidiot/431357888/

Data Capacitor as Central Filesystem



Early 2007 - 10 GigE Single Client Tests

977 MB/s between ORNL and IU
Using 10Gb TeraGrid connection
Identical Dell 2950 clients
- *2x dual 3.0 GHz Xeon*
- *4GB RAM*
- *1 Myricom Myri10G card*
  - *Ethernet mode*
- *Lustre 1.4.7.1*

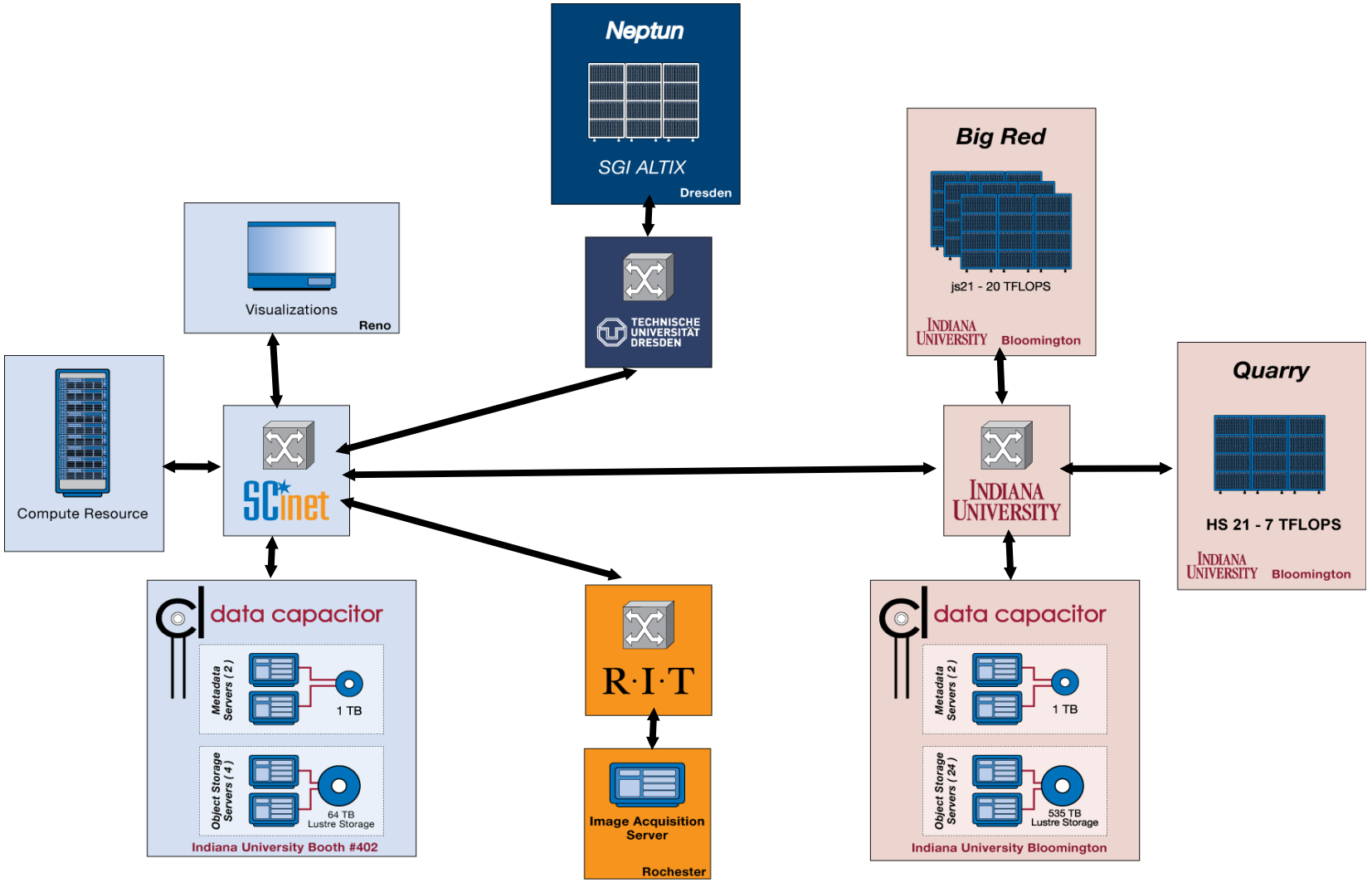
16 of 24 Data Capacitor (DC) OSSs

2007 Bandwidth Challenge Idea

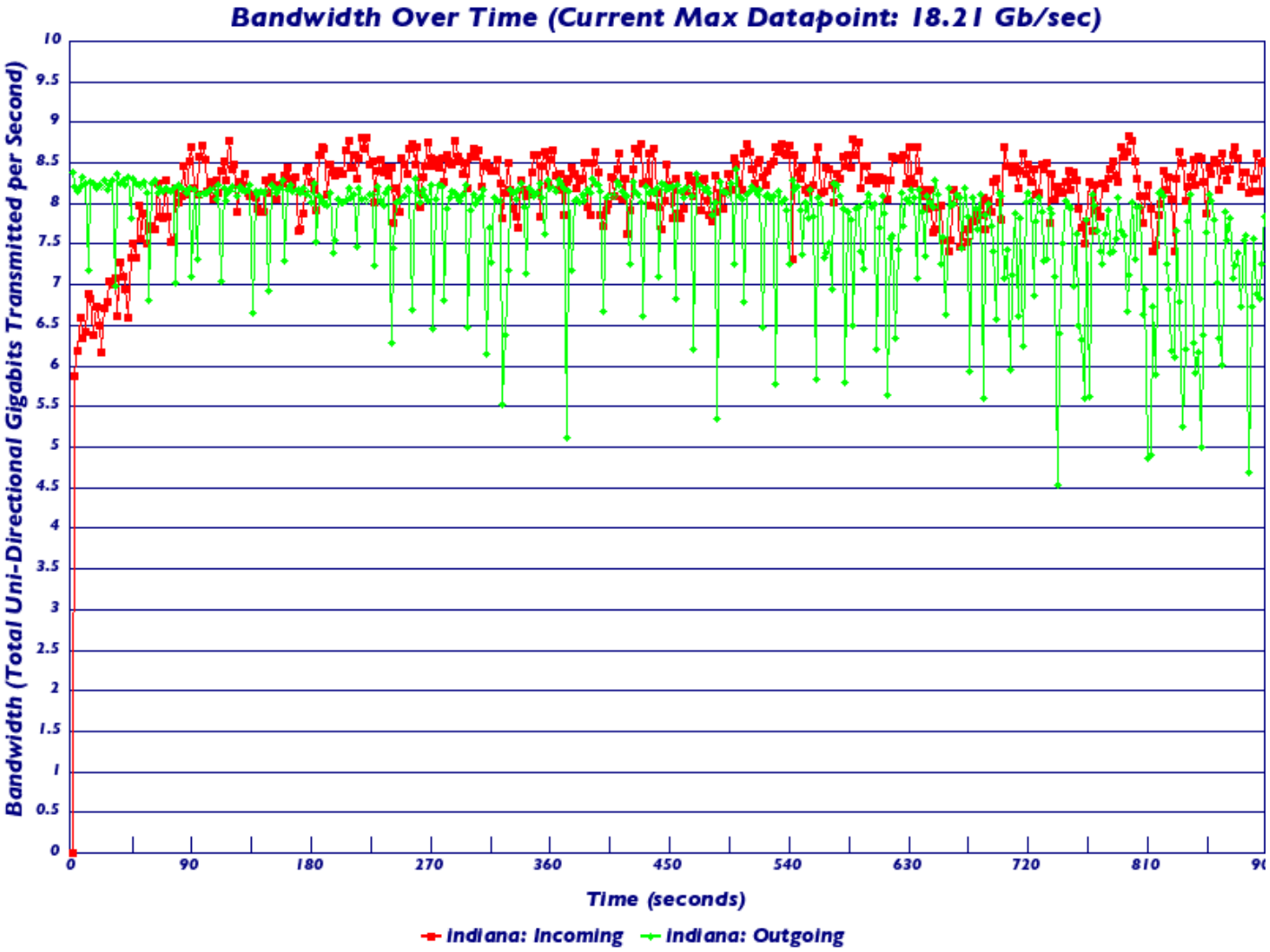
- Use the Data Capacitor across distance
  - DC supports over 100 Gigabit aggregate input
  - DC simple enough to build a small one in Reno
- Show science that UITS supports every day
  - We support a diverse community
- Use a production network
  - We used the Internet 2 network

INDIANA UNIVERSITY

Bandwidth Challenge Configuration



Bandwidth Challenge Results



Beyond a Demo – Namespace Mapping

- Challenge for Wide Area Filesystems is the creation of a homogeneous namespace for users across sites
- The numeric user identification (UID) for a particular user is not the same across sites

Indiana	TACC	PSC	NCSA	SDSC
jupmille	tg803934	jupmille	jupmille	jupmille
uid=648424	uid=803934	uid=43415	uid=40436	uid=502639

IU's UID Mapping for Lustre

Developed in 2007

Lightweight

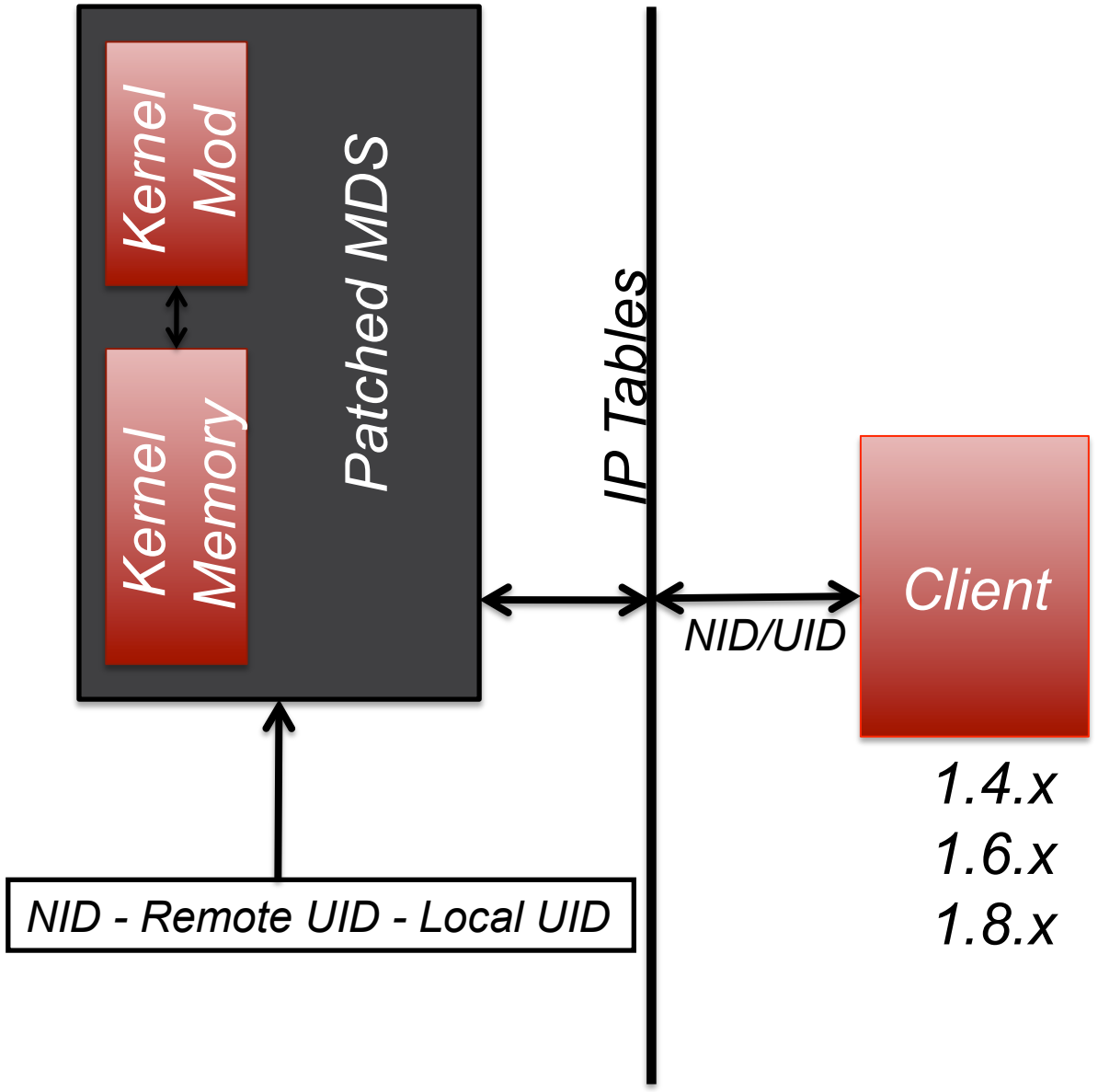
Changes made to the MDS code only

- Want to maximize clients we can serve

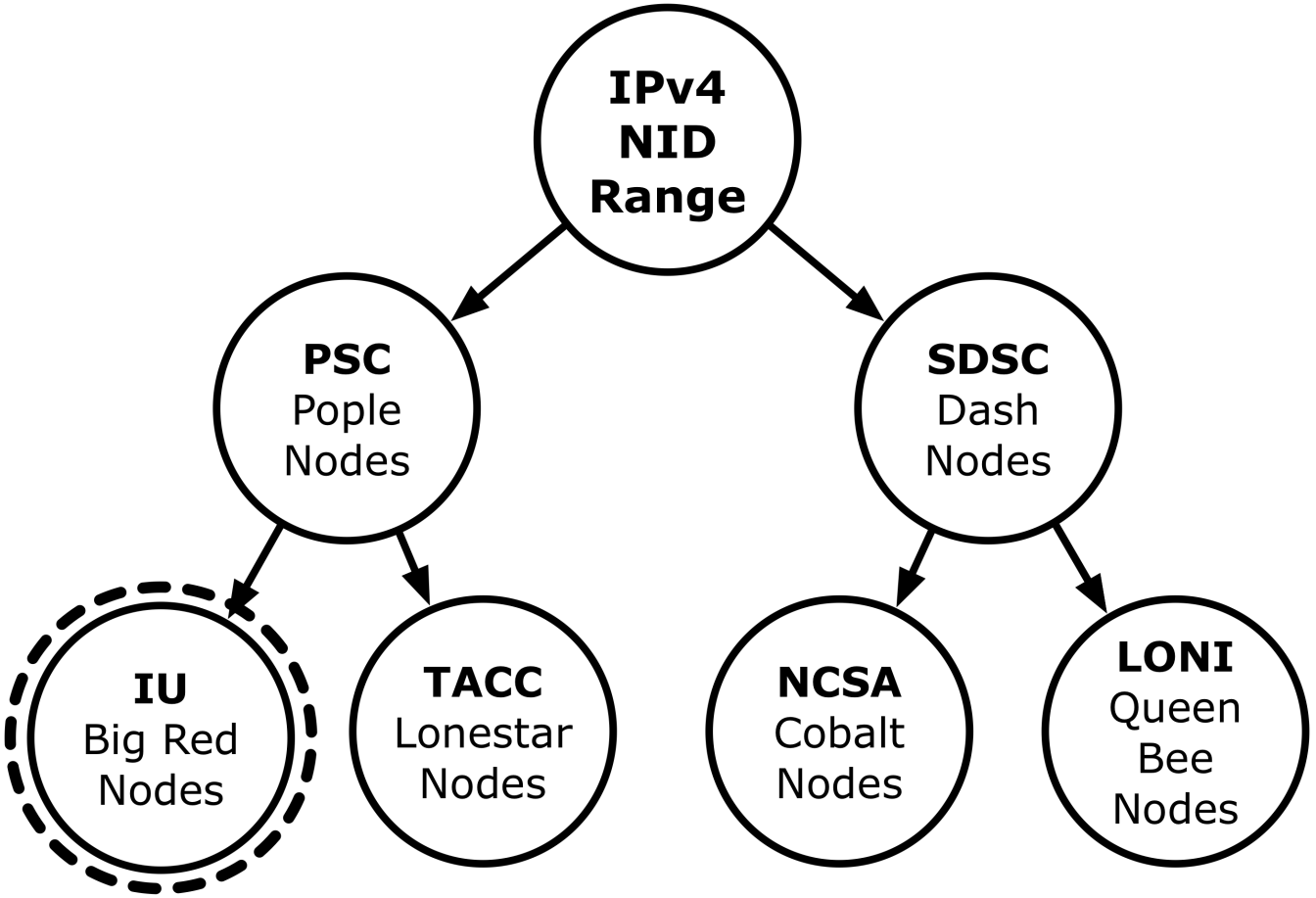
Customers can use standard clients

- Currently 1.4.x, 1.6.x, 1.8.x, unkerberized 2.0

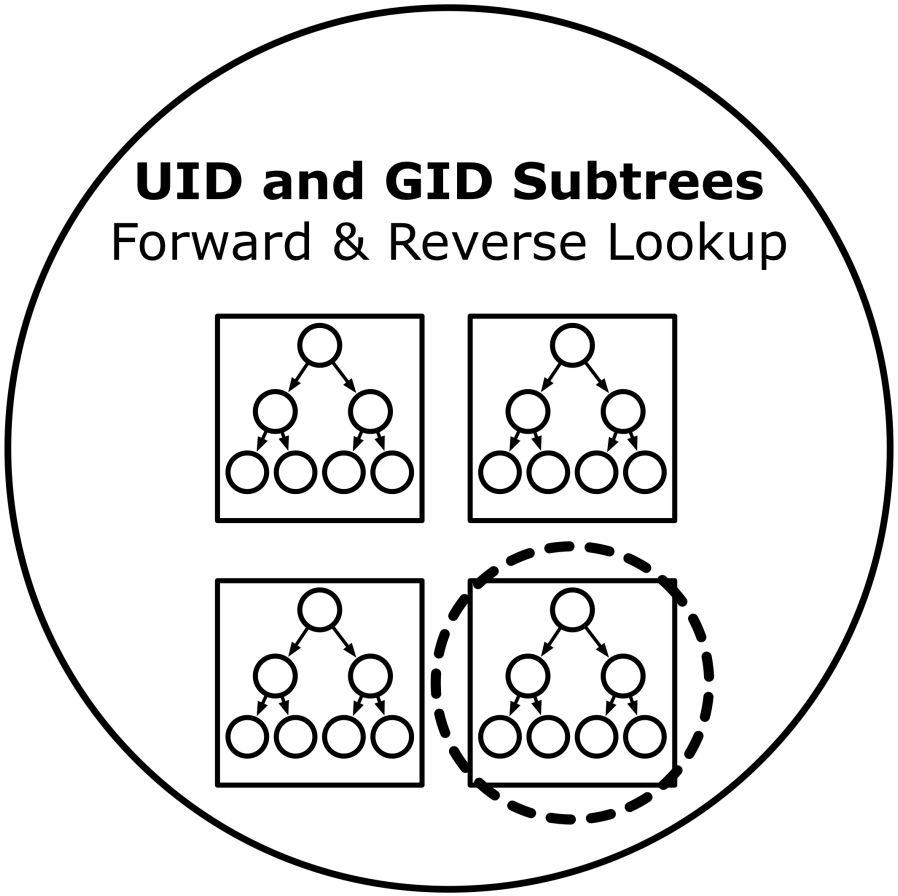
Until now simple enough to port the code forward


Kernel Mod
Kernel Memory
Patched MDS
IP Tables
NID/UID
Client
NID - Remote UID - Local UID
1.4.x
1.6.x
1.8.x

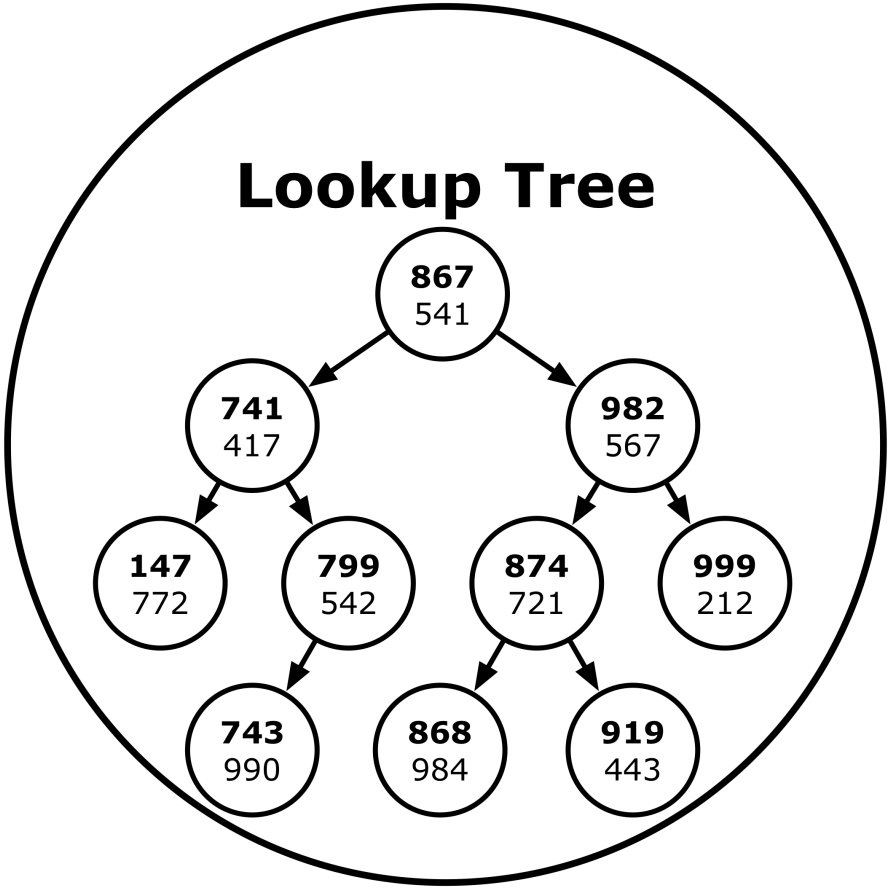
Forest of Binary Trees



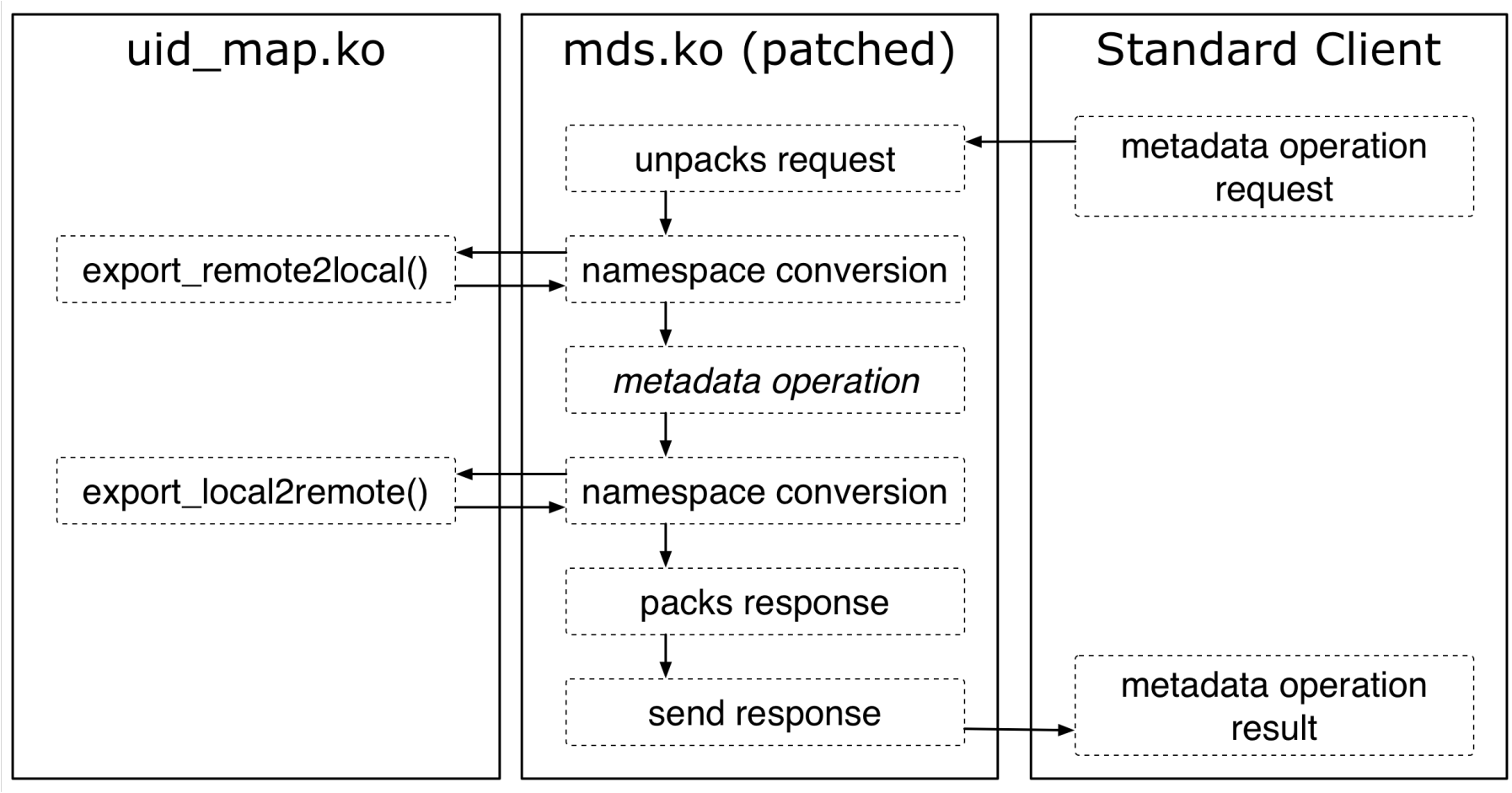
Forest of Binary Trees



Forest of Binary Trees



Patched MDS Transaction



IU's Data Capacitor WAN Filesystem

- Funded by Indiana University in 2008
- Put into production in April of 2008
- 340TB of storage available as production service
- Centralized short-term storage for nationwide resources, including TeraGrid

  Simplifies use of distributed resources

  Projects space exists for mid-term storage

ID Mapping in Lustre 2.x

Current UID mapping code:

1) does not support quotas
   quotas require UID continuity between MDS and OSS

2) relies on /proc interface
   only small messages between kernel and userspace
   loading large maps slow

3) will not support Lustre 2.X as a server

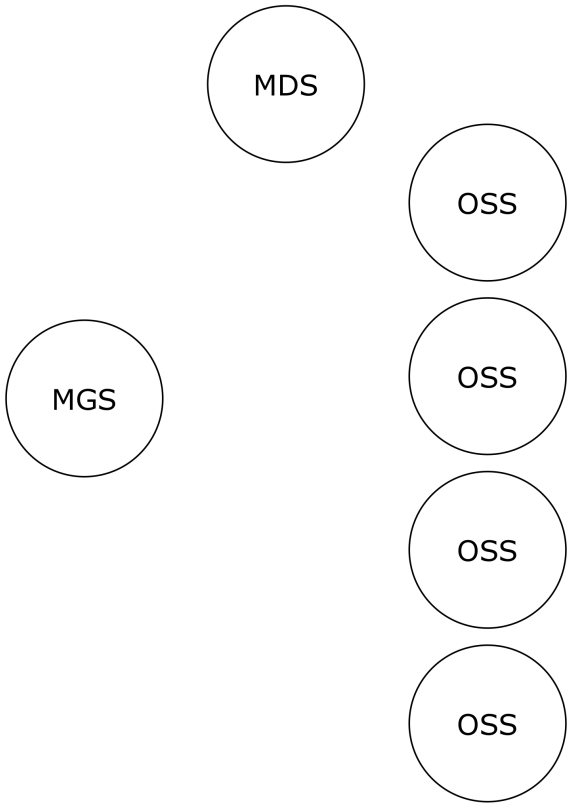
ID Mapping in Lustre 2.x

New UID mapping code will

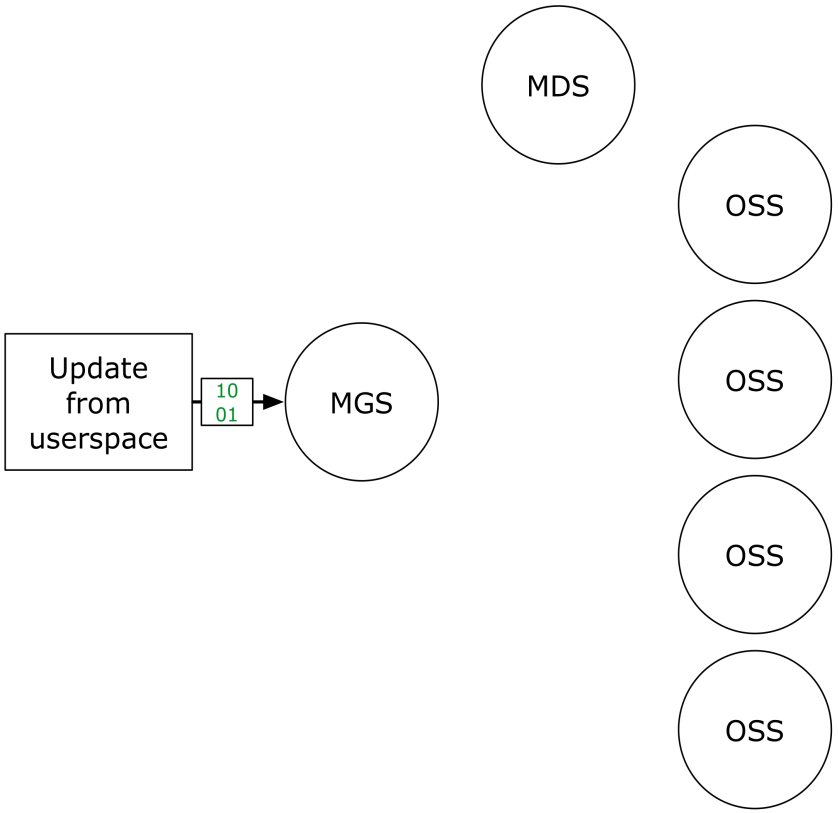
1) Live on the all Lustre servers
   - map controlled by MGS using ptlrpc
   - map updates will require synchronous locking

2) Provide userspace tools for map updates
   - netlink sockets are used for userspace configuration
   - interface with account management systems

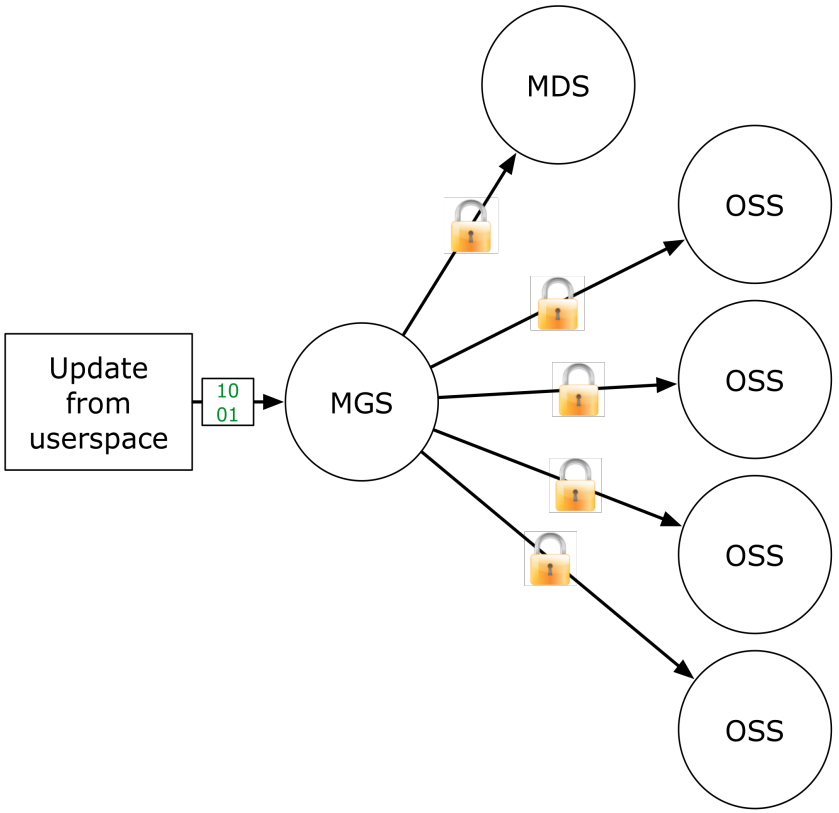
Normative State



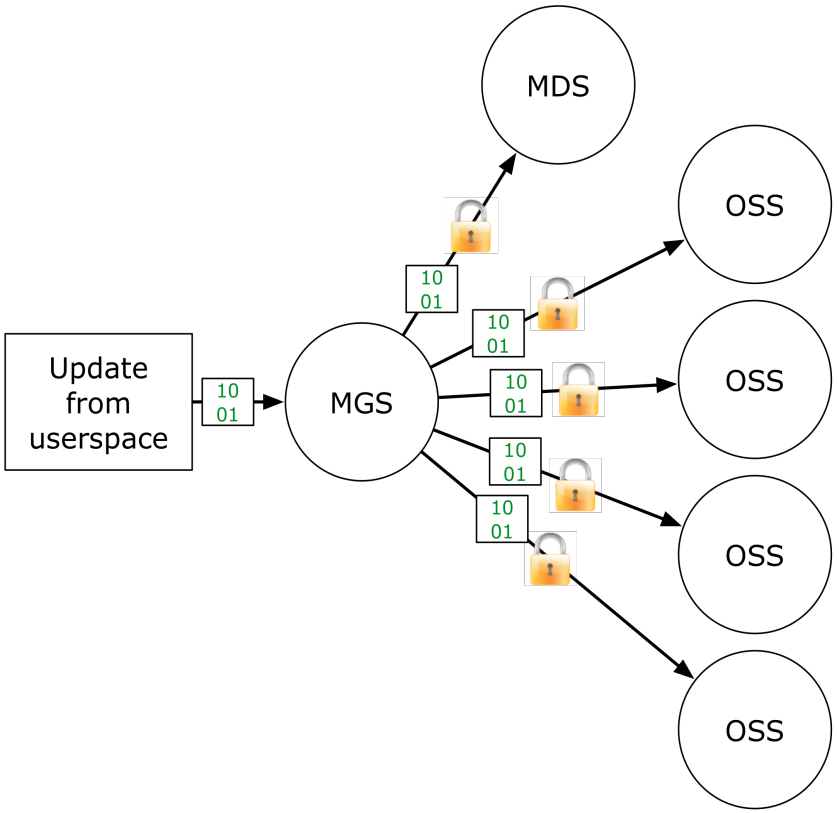
Admin Updates Map on MGS



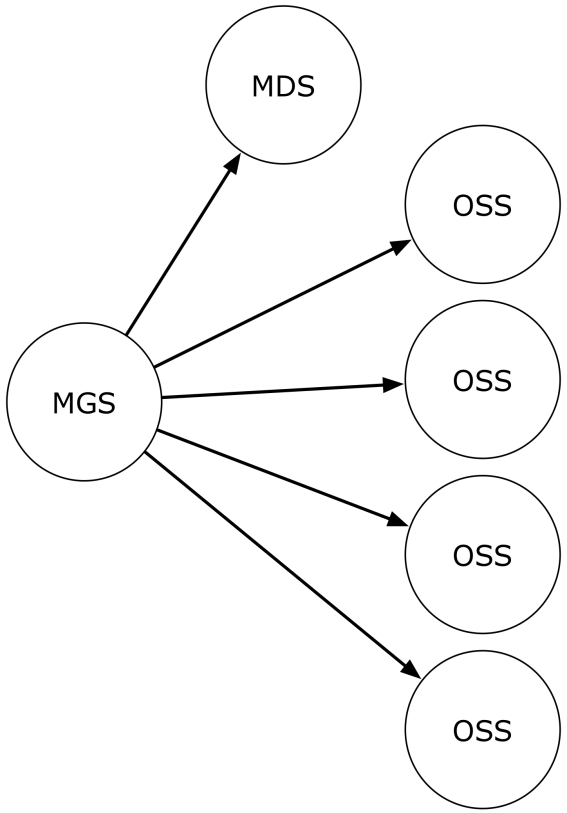
MGS Gets Lock on All Server Maps



Map Deltas Sent to Servers



Locks Removed Update Complete



Changes to Existing Lustre Codebase

Heavy lifting performed by additional kernel module which itself minimizes changes to the existing tree.

Changes to existing codebase limited to mapping the UID and GID after unpacking a request and packing the response.

Funding

- A proposal has been submitted to OpenSFS to fund this work.
- The OpenSFS Technical Working Group has reviewed this plan favorably and submitted it to the OpenSFS board.

Acknowledgements

- IU's High Performance File System Team
- Kit Westneat (NYU)
- Whamcloud

  Eric Barton, Andreas Dilger, Robert Read
- Craig Stewart, Matt Link (IU)
- The OpenSFS Technical Working Group

  John Carrier (Cray), Dave Dillow (ORNL)
- Data Direct Networks

This material is based upon work supported by the National Science Foundation under Grants No. CNS-0521433, ACI-0338618l, OCI-0451237, OCI-0535258, and OCI-0504075.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Thank You

Stephen Simms
ssimms@iu.edu

Josh Walgenbach
jjw@iu.edu